Probability Plots

Summary

The **Probability Plots** procedure plots the data in a single numeric column on graphs that are specifically scaled such that, if the data come from a particular distribution, the observations will fall approximately along a straight line. The procedure includes plots for the uniform, normal, lognormal, Weibull, smallest extreme value, logistic, and exponential distributions.

Sample StatFolio: *probplots.sgp*

Sample Data:

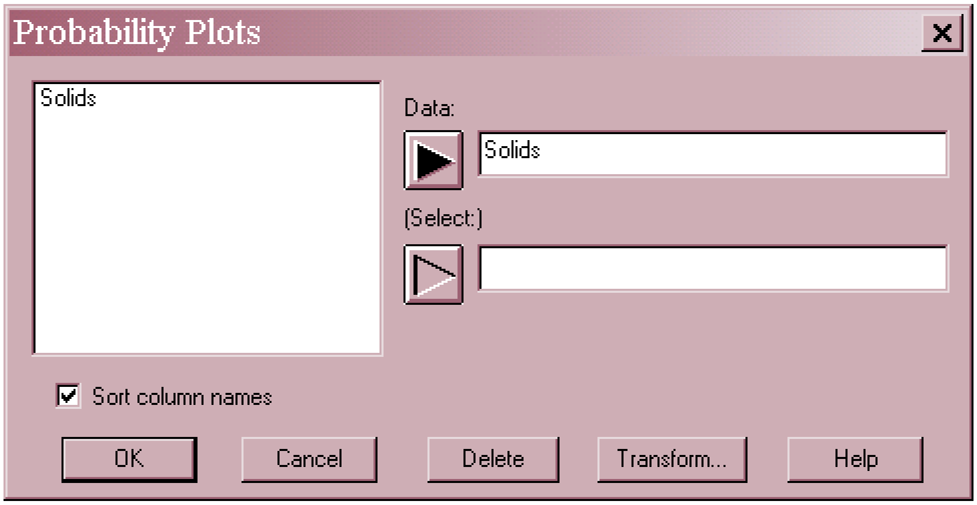
The file *groundwater.sgd* contains *n* = 47 measurements of the concentration of uranium in groundwater samples taken from a location in northwest Texas. The table below shows a partial list of the data from that file:

ppm
8.25
2.82
4.16
18.66
12.72
8.75
2.29
7.22
9.76
7.72
27.38
5.14

The concentration is measured in parts per million.

Data Input

The data to be analyzed consist of a single numeric column containing *n* = 2 or more observations.



Data: Solids

- **Data :** numeric column containing the data to be plotted.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows the number of observations plotted.

```
Probability Plots
Data variable: ppm
Number of observations = 47
Number of values below minimum: 0
Number of values above maximum: 0
```

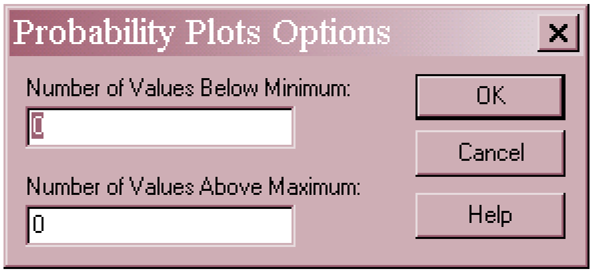
If the data are left-censored or right-censored, you may also use *Analysis Options* to indicate the number of observations that were taken (but not contained in the file) that are below the smallest value in the data column or above the largest value in the column. This is useful in cases such as:

1. When there is a lower limit of detection below which the measurement equipment reads zero, even though the actual value is not really zero.

2. In a life test where the study is stopped before all items have failed, so that the failure times of some items are not known exactly but are known to be greater than the time when the test was stopped.

For censored data, the plotting positions on the vertical axis of the probability plots is adjusted to compensate for the number of censored observations.

Analysis Options



Number of Values Below Minimum: 0

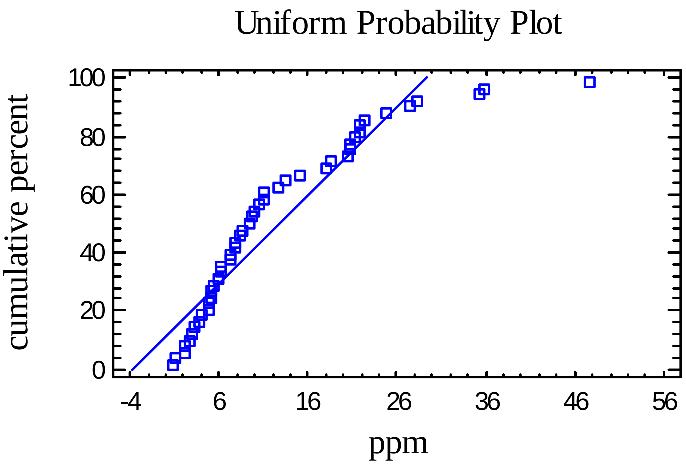
Number of Values Above Maximum: 0

- **Number of Values Below Minimum**: for left-censored data, the number of observations not contained in the data column that were less than the smallest value in the column.

- **Number of Values Above Maximum**: for right-censored data, the number of observations not contained in the data column that were greater than the largest value in the column.

Uniform Plot

The *Uniform Plot* is designed to help determine whether the data come from a continuous uniform distribution.



The data values are first sorted from smallest to largest and then plotted at the coordinates:

$$\left(x_{(j)}, r_j\right) \tag{1}$$

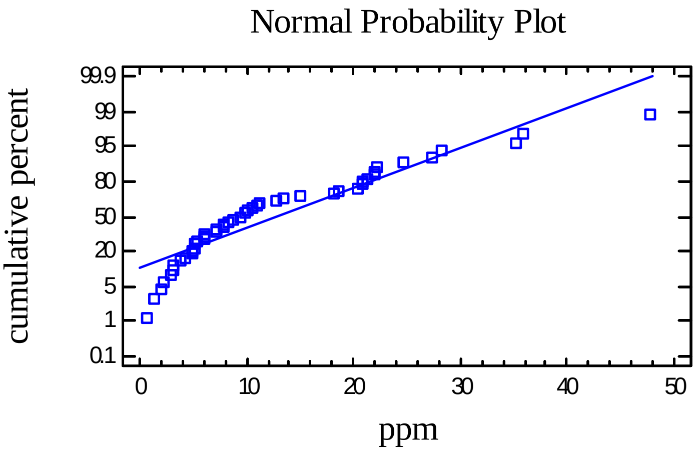
where r_j are called the median ranks, defined by

$$r_j = \frac{(j-0.3)}{n+0.4} \tag{2}$$

If the data values come from a uniform distribution, the points should fall approximately along the indicated line, which is fit by least squares regression of the data values against the median ranks. For the sample data, the plot shows noticeable curvature, indicating that the data are not well characterized by a uniform distribution.

Normal Plot

The *Normal Plot* is designed to help determine whether the data come from a normal distribution.



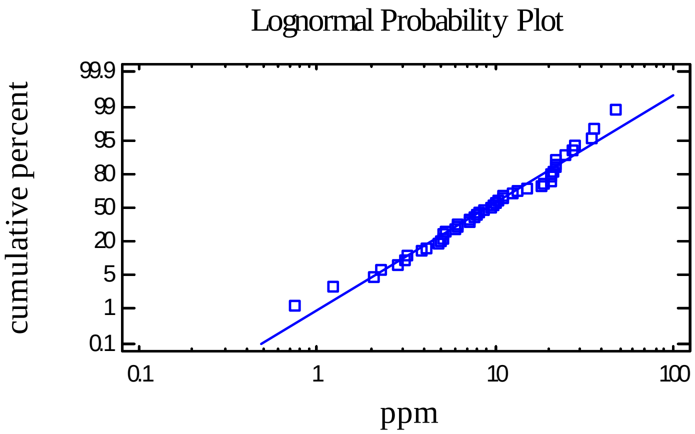
In this plot, the points are plotted at:

$$\left(x_{(j)}, \Phi^{-1}\left(\frac{j - 0.375}{n + 0.25} \right) \right) \tag{3}$$

where $\Phi^{-1}(u)$ represents the value of the inverse standard normal distribution evaluated at *u*. Compared to a normal distribution, the above sample has a short lower tail, i.e., the lower percentiles are not as far below the median as would be expected from a normal bell-shaped curve. This is often seen when there is a fixed lower bound such as *x* = 0.

Lognormal Plot

The *Lognormal Plot* is designed to help determine whether the data come from a 2-parameter lognormal distribution.



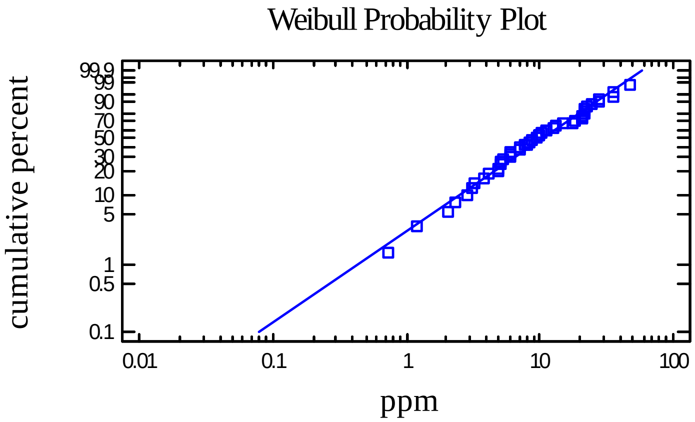
The coordinates of the points on this plot are:

$$\left(\log\left(x_{(j)}\right), \Phi^{-1}\left(\frac{j-0.375}{n+0.25}\right)\right) \tag{4}$$

where $\Phi^{-1}(u)$ again represents the value of the inverse standard normal distribution evaluated at u. There is a small amount of curvature above the line in the tails of the distribution, but the fit is far better than the earlier choices.

Weibull Plot

The *Weibull Plot* is designed to help determine whether the data come from a 2-parameter Weibull distribution.



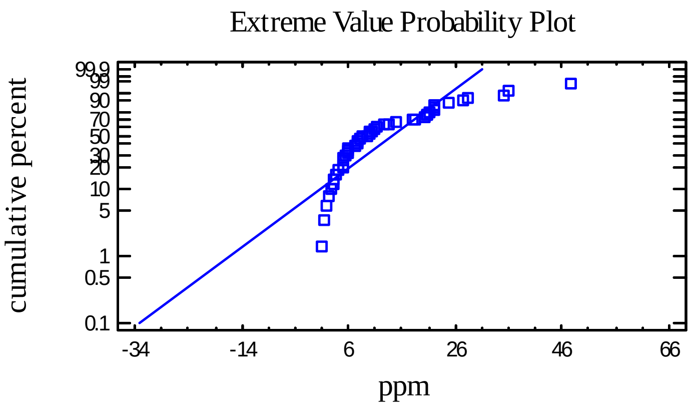
The coordinates of the points on this plot are:

$$\left(\log\left(x_{(j)}\right), \ln\left(-\ln\left(1-r_j\right)\right)\right) \tag{5}$$

In this case, the points lie very close to the line, indicating that the Weibull distribution would be a good candidate for this data.

Extreme Value Plot

The *Extreme Value Plot* is designed to help determine whether the data come from a smallest extreme value distribution.



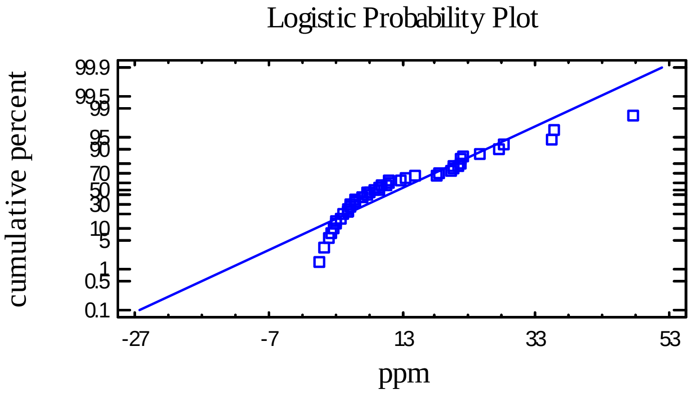
The points are plotted at:

$$\left(x_{(j)}, \ln\left(-\ln\left(1-r_j\right)\right)\right) \tag{6}$$

The smallest extreme value distribution is negatively skewed, which is not at all appropriate for the current data. This distribution is often used for log failure times, however, since it forms the proper model for the logarithm of a random variable which itself follows a Weibull distribution.

Logistic Plot

The *Logistic Plot* is designed to help determine whether the data come from a logistic distribution.



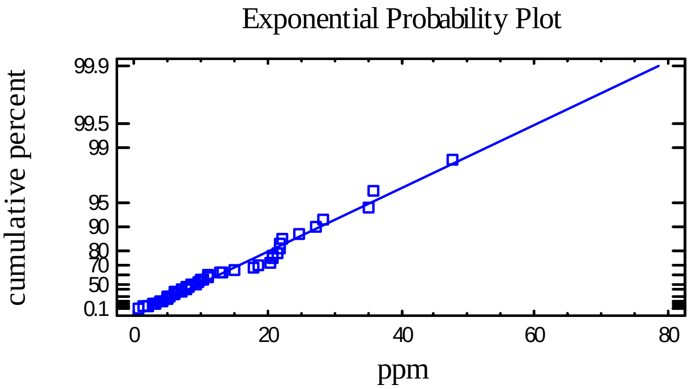
In this plot, the points are plotted at:

$$\left(x_{(j)}, \frac{\sqrt{3}}{\pi}\ln\left(\frac{r_j}{1-r_j}\right)\right) \tag{7}$$

Again, the logistic distribution does not match the data well.

Exponential Plot

The *Exponential Plot* is designed to help determine whether the data come from an exponential distribution.



The points are plotted at:

$$\left(x_{(j)}, -\ln\left(1-r_j\right)\right) \tag{8}$$

The exponential distribution, which peaks at *x*=0, is also a poor choice for the current data sample.